

# Probabilistic Topic and Syntax Modeling with Part-of-Speech LDA

William Darling\*  
Xerox Research Centre Europe

Fei Song\*\*  
University of Guelph

*This article presents a probabilistic generative model for text based on semantic topics and syntactic classes called Part-of-Speech LDA (POSLDA). POSLDA simultaneously uncovers short-range syntactic patterns (syntax) and long-range semantic patterns (topics) that exist in document collections. This results in word distributions that are specific to both topics (sports, education, ...) and parts-of-speech (nouns, verbs, ...). For example, multinomial distributions over words are uncovered that can be understood as “nouns about weather” or “verbs about law”. We describe the model and an approximate inference algorithm and then demonstrate the quality of the learned topics both qualitatively and quantitatively. Then, we discuss an NLP application where the output of POSLDA can lead to strong improvements in quality: unsupervised part-of-speech tagging. We describe algorithms for this task that make use of POSLDA-learned distributions that result in improved performance beyond the state of the art.*

## 1. Introduction

Two highly related phenomena are resulting in a renaissance for the study of computational linguistics. The first is the increasing level of access to textual data sources over the Internet such as classic works of literature (The Gutenberg Project), structured explanatory knowledge of the world (Wikipedia), people’s every thought (Twitter), governments’ every desire (laws and legal decisions), and ongoing triumphs, defeats, changes, and breaking information (online news). The second, concomitant with the first, is the growth of interest in, and the power of, machine learning algorithms which can exploit the vast amounts of data that are being made available and help make sense of them.

Like language itself, machine learning techniques can be described and contrasted with each other along a number of different axes of understanding and dichotomies. One of the most important of these dichotomies is the division between *supervised* and *unsupervised* learning approaches (Blei, Griffiths, and Jordan 2010). While supervised approaches are concerned with generalizing a function that predicts an output  $y$  given some input  $x$  learned from examining labeled example pairs  $(x_i, y_i)$ , unsupervised approaches involve uncovering hidden patterns and associations that exist in data (Hastie, Tibshirani, and Friedman 2001). Clustering algorithms, for example, are unsupervised machine learning techniques that attempt to group data together because they are similar in some way. A logical definition of similarity – especially with respect to linguistics – is, informally, that two texts are similar because they discuss the same topics.

---

\* 6 chemin de Maupertuis, 38240 Meylan, France. E-mail: william.darling@xrce.xerox.com.

\*\* 50 Stone Road East, Guelph, Ontario, N1G 2W1, Canada. E-mail: fsong@uoguelph.ca.

Another canonical example of unsupervised learning is dimensionality reduction. In reducing the dimensionality of a dataset, an algorithm seeks to find a simpler or more condensed representation of the data while preserving (or bringing forth) some kind of meaning. An apt dimensionally-reduced representation of texts – collections of words – is the topics that they represent. In the standard bag-of-words representation used for many natural language processing tasks, the cardinality of the dimensional representation is the number of distinct words that can be used in the texts, that is, the vocabulary. This number is typically on the order of several thousand, while a text's content, in a broad sense, can also be described by the topics that it addresses out of a finite number of generic topics on the order of hundreds or less.

It is therefore no surprise that probabilistic topic models, which are generative models of text (based on unsupervised machine learning algorithms), are continually growing in interest. They provide a means to uncover the hidden thematic structure that underlies large document collections and therefore allow us to explore, summarize, and understand what a collection is about with ease and efficiency. Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), the original topic model, describes the generative story that lays the foundation for this kind of model. It posits that a document can be created through a random process of drawing words from a mixed-membership model. At a high level, a document is created by first selecting the topics that it will address, and then randomly generating words from distributions associated with those topics. Numerous more complex models have been presented that build on this basic idea and several introductory papers exist that cover the area in detail (Blei 2012).

More specifically, the LDA generative process works as follows. For each document  $d$ , a document-specific topic portion  $\theta_d$  is drawn from a Dirichlet distribution.  $\theta_d$  is a discrete distribution over  $K$  topics and corresponds to the weight that each topic will have in the document. Then, for each word  $w_i$ , a topic index  $z_i$  is drawn from  $\theta_d$ . To generate the word, a token is drawn from a topic-specific word distribution  $\phi^{(z_i)}$ . There are  $K$  topic-specific word distributions, each of which corresponds to a distribution over words specific to the given topic. To generate a document, however, one would require the word distributions specific to each topic, and for each document, one would also need the topic portion. Neither of these are readily available from the input, but they can be learned from a document collection by reversing the generative process through posterior inference.

In unsupervised learning, we cannot simply write a general algorithm to find what we hope to be interesting patterns. The “no free lunch” theorem tells us essentially that no interesting patterns can be uncovered if we do not assume that certain kinds of patterns must exist (Wolpert and Macready 1997). It turns out that assuming that documents reflect latent topics and that words from the same topics will co-occur is a good assumption and thus interesting results can be learned through reversing the assumed model such as the words that are most important to each topic, and a dimensionally-reduced representation of each document in topic space. However, the correct number – or the type of assumptions – is important. If we make too many, the data may not fit those assumptions and the output will be nonsense. Conversely, if we make too few, interesting patterns may be missed.

The standard document representation for topic modeling is the bag-of-words (Salton and McGill 1983). Each document is represented by the number of times that each word in a fixed vocabulary appears. Because word order is ignored, a great deal of meaning is lost, but the representation is efficient and has proved to be successful. However, word order – ignored in canonical topic models – is clearly of importance to

language because though some sense may be extracted from a text whose words have been scrambled, the full meaning can only come about through a parsing of the words based on their order and relation to each other. It has also been shown that different parts of the brain are used to understand semantics and syntax (Boyd-Graber and Blei 2010). Traditional topic models miss this kind of syntactical information because they are never exposed to word-order patterns in the first place. While the bag-of-words approach is efficient, further advances in NLP will require algorithms to be able to have the full understanding that humans are afforded.

In this article we introduce a new probabilistic generative model called Part-of-Speech LDA (POSLDA) which considers both the semantic topic that a word is associated with (if any) and its syntactic purpose in a sentence. This approach allows a more structured view of language creation and as a result, the posterior distributions that are learned are more specific and meaningful than in other topic models. It also allows NLP applications to make better predictions and deductions because each piece of information – syntactic and semantic – provides evidence that can help disambiguate the purpose and meaning of a word.

The article is organized as follows. In section 2, we discuss previous work that has focused on bringing syntactic information into probabilistic topic models. In section 3, we explain our model, POSLDA, in detail. We then present the results of three sets of experiments in section 4: first, we demonstrate qualitatively the interpretability of the uncovered posterior distributions with example syntax-specific topics on a number of diverse datasets; second, we report quantitative results on the model’s ability to generalize on unseen texts and to uncover high quality topics; and third, we show how the model’s ability to disambiguate word use through the joint influences of semantics and syntax can lead to better results in unsupervised part-of-speech (POS) tagging than a Bayesian Hidden Markov Model (HMM). Finally, in section 5, we conclude with thoughts on future work.

## 2. Topics and Syntax

The constraints that are imposed by language on phrase structure and word order are called syntax (Manning and Schütze 1999). The syntactic meaning of a word helps to explain its functional purpose in a sentence, whereas the semantic meaning is related to its lexical-thematic purpose. The former is based on short-range dependencies at the sentence level, while the latter realizes long-range dependencies at the document level. LDA and other topic models uncover patterns by exploiting the long-range dependencies of words co-occurring. Here, we want to add the short-range dependency structures to the model and we do so by focusing on modeling the *functional* purpose of a word in a sentence. We are therefore interested in the part-of-speech category that a word – in its given context – belongs to. These include nouns, verbs, adjectives, adverbs, prepositions, conjunctions, etc. The canonical tool for unsupervised word syntax modeling is the hidden Markov model (HMM) (Rabiner 1990) and it is therefore a natural place to begin in adding syntax information to a semantic topic model.

The first work in combining syntactic notions of language with probabilistic topic models is based on embedding an LDA-like model in a single state of an HMM (Griffiths et al. 2005). This model – dubbed HMMLDA – represents an asymmetric composite model where all generated words follow short-range syntactic dependencies, but only “semantic” words that are generated from a single state obey long-range dependencies. Words that carry long-range dependencies will be generated given the document-specific topic distribution, and other words will be generated independent of the current

theme. This framework is different from the traditional use of the HMM in syntax modeling as each state will not correspond to a discrete part-of-speech. The content class in particular – which is designated as the sole class that can generate “semantic” words – will need to subsume nouns, verbs, adjectives, and other words that are topic-dependent. We will return to this simplifying assumption when we present our POSLDA model.

More formally, the HMMLDA model is defined by two sets of sequential latent variables  $(z_n)_{n=1}^N$ , which represent the latent topics for each word, and  $(c_n)_{n=1}^N$ , which represent the latent syntactic classes for these words. One state in the model,  $s_0 \in S$ , is designated as the semantic class where the LDA-like topic model is embedded. Each topic  $k$  is associated with a discrete topic-word distribution  $\phi^{(k)}$  and each class  $s \neq s_0$  is associated with a syntax word distribution  $\phi^{(s)}$ . Like LDA, each document  $d$  can be described by a distribution over topics  $\theta^{(d)}$ . However, unlike LDA, each word  $w_i$  only depends on its topic  $z_i$  if its class  $c_i = s_0$ . A word’s class is modeled with the embedded HMM and transitions between classes  $s_i$  and  $s_{i+1}$  are encoded in a transition matrix  $\pi$ . The generative process by which a document is created under the HMMLDA model is as follows.

1. Draw  $\theta^{(d)} \sim \text{Dirichlet}(\alpha)$
2. For each word  $w_i$  in document  $d$ 
  - (a) Draw topic  $z_i \sim \theta^{(d)}$
  - (b) Draw class  $c_i \sim \pi^{(c_{i-1})}$
  - (c) If  $c_i = s_0$ :
    - i. Draw  $w_i \sim \phi^{(z_i)}$
  - (d) Else:
    - i. Draw  $w_i \sim \phi^{(c_i)}$

Like LDA, exact posterior inference is intractable for the HMMLDA model (Griffiths et al. 2005). Griffiths, et al. therefore turn to Gibbs sampling. The HMM in the HMMLDA is a Bayesian HMM meaning that the transition rows  $\pi_r$  and the emission probabilities  $\phi^{(c)}$  are multinomial random variables with Dirichlet priors. The same framework for collapsed Gibbs sampling can therefore be used as with the original LDA.

One of the most interesting qualities of the HMMLDA model is that stop-words and other “syntax-only” words are pulled to the non-semantic classes so that the learned topics are interpretable and noise-free without any need for pre-processing or *a priori* stop-word removal. This defines one of the key motivations in the development of POSLDA. While the model seems to learn distributions that are noticeably useful, it is still far from perfect. It almost exclusively finds nouns as content words when verbs are often equally as important in semantic topics. This could be due to the use of the HMM that learns to discern nouns from other parts-of-speech and simply sees other semantically-important types of words as outliers that are pushed to the syntactic classes.

Another recent approach at combining syntax and semantics into a coherent probabilistic generative model is the Syntactic Topic Model (STM) (Boyd-Graber and Blei 2010). Unlike the HMMLDA model, where a word is deemed to either come from a corpus-wide syntax class *or* a semantic-based topic, the STM discovers topics that are both syntactically and semantically coherent. This is more directly in line with our goals for the POSLDA model. As for the motivation in combining these notions of word information, Boyd-Graber and Blei provide an edifying example:

**Table 1**  
Example Topics from the STM

hates	bucks	runs	professor	stock
dreads	surges	falls	phd	share
mourns	climbs	walks	candidate	mutual
fears	falls	sits	grad	fund
despairs	runs	climbs	student	on

Next weekend, you could be relaxing in \_\_\_\_\_.

There are two distinct text-modeling based approaches one could use to reason about what word might be used to fill in the blank. With a topic model such as LDA, where this document is about travel, high probability words might include “sailing”, “Rome”, or “flight”. With a syntax model, it might determine that the missing word should be a noun, and thus high probability words could include “bed”, “church”, or “school”. However, as is explained in (Boyd-Graber and Blei 2010), the best candidate to fill in the blank is an intersection of these two types of reasoning. The word “sailing” matches the topic related aspect of the sentence (travel), but does not fit syntactically (verb). The opposite is the case for the noun “school”. However, when both syntactic and semantic notions of language are taken into account, a word such as “Rome”, which has high probability both in the travel topic and in the noun syntax class, will be selected with high probability.

The STM is a non-parametric model where the number of topics is not set *a priori* but is determined, through posterior inference, by the data. While the standard LDA model draws the document topic portions for a document from a  $K$ -dimensional Dirichlet distribution with a fixed value of  $K$  topics, here the transition distributions  $\pi$  and document topic portions  $\theta_d$  are drawn from a Dirichlet Process (DP), where a vector  $\beta$  of infinite length is a global weight that is drawn from a stick-breaking distribution and is used as a base measure for the DP. This approach frees users of the model from having to determine themselves how many topics a corpus might contain. Our POSLDA model can also easily become a nonparametric Bayesian model by incorporating a hierarchical Dirichlet Process (HDP) prior (Teh et al. 2006).<sup>1</sup> In this formulation, the number of topics can be learned from the data in the sense that through inference we can learn the optimal number of topics  $K$  that will maximize the likelihood of the data.

The generative process for the STM is as follows:

1. Draw global weights  $\beta \sim \text{GEM}(\alpha)$
2. For each topic index  $k = \{1, \dots\}$ :
  - (a) Draw topic  $\tau_k \sim \text{Dir}(\sigma \rho_u)$
  - (b) Draw transition distribution  $\pi_k \sim \text{DP}(\alpha_T, \beta)$
3. For each document  $d = \{1, \dots, M\}$ :
  - (a) Draw document weights  $\theta_d \sim \text{DP}(\alpha_D, \beta)$
  - (b) For each sentence root node with index  $(d, r) \in \text{SENTENCE-ROOTS}_d$ :
    - i. Draw topic assignment  $z_{d,r} \propto \theta_d \pi_{start}$
    - ii. Draw root word  $w_{d,r} \sim \tau_{z_r}$

<sup>1</sup> The approach to do so specifically for POSLDA is outlined in §3.4 of (Darling 2012).

- (c) For each additional child with index  $(d, c)$  and parent with index  $(d, p)$ :
  - i. Draw topic assignment  $z_{d,c} \propto \theta_d \pi_{z_{d,p}}$
  - ii. Draw word  $w_{d,c} \sim \tau_{z_{d,c}}$

While this generative process is clearly much more complex than the corresponding one for simpler models such as LDA and HMMLDA, it is also more powerful. The key steps are 3(b)(i) and 3(c)(i) where the topic assignment for a word is chosen to be a convolution of the long-range semantic topic portion  $\theta_d$  and the short-range syntactic probability  $\pi_{z_{d,p}}$ . Some example semantically and syntactically coherent topics learned from the STM are shown in Table 1. Note that in each case, the high probability words are both syntactically equivalent (noun, verb, etc.) and semantically related in a topic modeling sense.

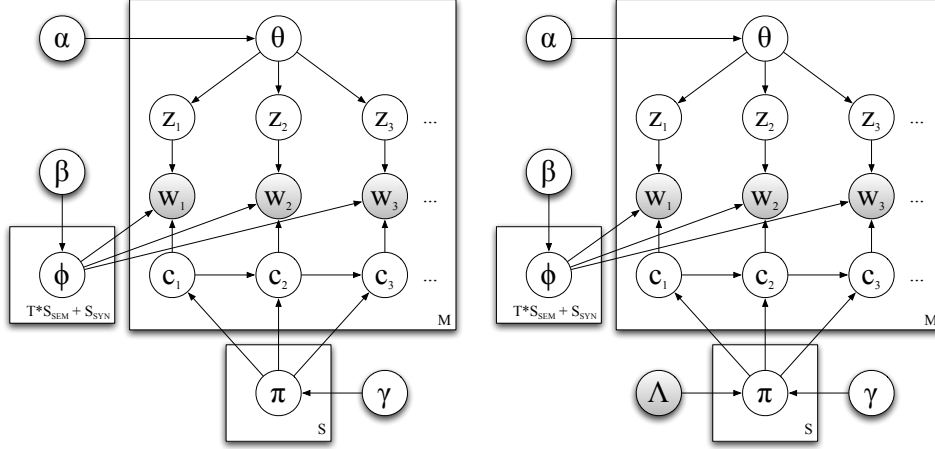
While the posterior distributions learned by the STM appear to fall in line with our goals of a syntactically-cognizant generative topic model, it suffers from certain limitations that we would like to address. First, the generative story depends on a form of meta-sentence structure that exists before the words have been generated. That is, texts are generated based on the probabilities in sentence-specific dependency parse trees (Boyd-Graber and Blei 2010). Therefore, to perform inference with the STM – and recover the posterior distributions like those in Table 1 – the data must be separately pre-processed into sentence dependency parses. This means that the model is not learning the syntax patterns itself, but is using information that must be supplied by a separate algorithm before inference is performed. Conversely, we are interested in a fully generative model that is consistent across syntax and semantics where inferring the short-range sentence-wide dependencies forms part of the model.

In the next section we will present our model, POSLDA. It is a strong generalization of HMMLDA in that it follows the idea that we can combine an HMM with an LDA-like topic model, but it takes the idea further so that topic-dependent words can also be influenced by different syntax classes and thus learn part-of-speech specific posterior distributions.

### 3. POSLDA

While LDA is in some sense a simple extension of probabilistic latent semantic analysis (Hofmann 2001), it can be seen as the first fully generative topic model by virtue of the Dirichlet prior that is placed on the document-topic portions which in effect free the model from specific training data. Since its inception, LDA has been extended in numerous ways and particularly by infusing the model with additional factors. Word distributions can become more specific if we consider that generated words are dependent on not only the current topic, but also other latent aspects such as the sentiment of the writing and the writer’s personality or ideological perspective (Paul and Girju 2010; Ahmed and Xing 2010). This allows one to uncover such word distributions as “positive/negative words about films” or “words about weather from the perspective of Americans/Swedes/Australians”. In fact, this approach is so powerful that it has been generalized into techniques that can easily add specific factors to topic models through the use of strong prior information (Paul and Dredze 2012). However, to include word syntax, we need a different approach because this factor does not come out of the *types* of words that are used, but their *order*.

Part-of-Speech LDA (POSLDA) is an extension and generalization of LDA and HMMLDA that is designed to understand the long- and short-range dependencies between words, and as a tool for more complex NLP tasks that require both seman-

**Figure 1**

Graphical model depiction of POSLDA (left) and Labeled POSLDA (right) for unsupervised POS tagging.

tic and syntactic information to attain optimal performance. In an HMM, words are considered independent of their wider context within a document, but depend on the classes of the words that appear before them. Therefore, the word order in a syntax model is important and the bag-of-words representation used in canonical topic models is no longer appropriate. Because both types of word information are important, and modeling each separately entails certain restrictions, we seek to bridge these restrictions with a unified model of language, POSLDA.

Under POSLDA, each word token is now associated with two latent variables: a topic  $z$  and a syntactic class  $c$ . We posit that the topics are generated through the LDA process, while the classes are generated through a Bayesian HMM. The observed word tokens are then generally dependent on both the topic and the class: rather than a single discrete distribution for a particular topic  $z$  or a particular class  $c$ , there are distributions for each topic-class pair  $(z, c)$  from which we assume words are sampled. However, there also exists a set of “syntactic-only” words that do not depend on the thematic context of a document (Griffiths et al. 2005). These words – such as determiners, prepositions, and conjunctions – are often called “function” words and should be modeled as “universal” syntax classes that are not affected by – and are not assigned – a latent topic.

We therefore are interested in a generative model of text where all generated words depend on the function that they perform in a sentence, and a subset of these words also depend on the current semantic topic. For the HMM-like portion of the model, we denote the set of classes  $\mathcal{C} = \mathcal{C}_{\text{SEM}} \cup \mathcal{C}_{\text{SYN}}$ , which includes the set of semantic classes  $\mathcal{C}_{\text{SEM}}$  and the set of syntactic (function word) classes  $\mathcal{C}_{\text{SYN}}$ . If a word is generated from a function word class, it does not depend on the topic. This allows our model to accommodate functional words that appear independently of the topical content of a document.

We use a similar notation to LDA, where  $\theta_d$  is a document-topic portion and  $\phi_\eta$  is a word distribution. Additionally, we denote the HMM transition matrix  $\pi$ , which we assume has rows that are drawn from a Dirichlet distribution with hyperparameter  $\gamma$ . Denote  $S = |\mathcal{C}|$  and  $T = |\mathcal{Z}|$ , the numbers of classes and topics, respectively. There are

$S_{\text{SYN}}$  word distributions  $\phi^{(\text{SYN})}$  for function word classes and  $T \times S_{\text{SEM}}$  word distributions  $\phi^{(\text{SEM})}$  for semantic classes. A graphical model depiction of POSLDA is shown in Figure 1. This figure also denotes a slight variation to the model called Labeled POSLDA which is analogous to Labeled LDA for the original LDA (Ramage et al. 2009). Here, an observed dictionary denoted by  $\Lambda$  restricts the classes that certain words can take on. We will make use of this model for dictionary-based unsupervised part-of-speech tagging.

The generative process for a corpus of documents  $\mathcal{D}$  under the POSLDA model is described as follows:

1. For each row  $\pi_r \in \pi$ :
  - (a) Draw  $\pi_r \sim \text{Dirichlet}(\gamma)$
2. For each word distribution  $\phi_\eta \in \phi$ :
  - (a) Draw  $\phi_\eta \sim \text{Dirichlet}(\beta)$
3. For each document  $d \in \mathcal{D}$ :
  - (a) Draw  $\theta_d \sim \text{Dirichlet}(\alpha)$
  - (b) For each word token  $w_i \in d$ :
    - i. Draw  $c_i \sim \pi_{c_{i-1}, \dots, c_{i-n}}$
    - ii. If  $c_i \notin \mathcal{C}_{\text{SEM}}$ :
      - A. Draw  $w_i \sim \phi_{c_i}^{(\text{SYN})}$
    - iii. Else:
      - A. Draw  $z_i \sim \theta_d$
      - B. Draw  $w_i \sim \phi_{c_i, z_i}^{(\text{SEM})}$

In traditional topic models, it is generally the case that common function words will overwhelm the word distributions, leading to suboptimal results and learned word distributions that are difficult to interpret. This problem is often skirted by either data pre-processing (e.g. removing stop words from a domain-dependent list) (Blei, Ng, and Jordan 2003), backing off to “background” word models (Chemudugunta, Smyth, and Steyvers 2006; Paul and Girju 2010), or by performing term re-weighting (Wilson and Chew 2010). In the case of POSLDA, these common words are naturally explained by the corresponding function word classes and are pushed to these distributions rather than the topic-specific distributions during learning.

### 3.1 Relations to Other Models

The idea of having discrete word distributions for the cross product of topics and classes is related to multi-faceted topic models where word tokens are associated with multiple latent variables (Paul and Girju 2010; Ahmed and Xing 2010; Paul and Dredze 2012). Under such models, words can be explained by a latent topic as well as a second (or  $n$ th) underlying variable such as the perspective or dialect of the author, and words may depend on both (or multiple) factors. In our case, the second variable is the part-of-speech – or functional purpose – of the token.

POSLDA is also similar to a recent model called Nested HMM-LDA (nHMLDA) (Jiang 2009). The model described is very similar to POSLDA but contains certain limitations. Principally, rather than allowing each word to be generated from any of  $K$  topics, all words from a sentence must share the same topic. This is a strong assumption since it will not allow a sentence to discuss more than a single topic.

POSLDA is constructed in a generalized manner and contains many existing models as special cases. For example, POSLDA reduces to a Bayesian HMM when the



number of topics  $K = 1$ , the original LDA model when the number of classes  $S = 1$ , or the HMMLDA model when the number of semantic classes  $S_{\text{SEM}} = 1$ . One of the key benefits of these reductions is that one can easily experiment with any of these models using a single POSLDA implementation by simply altering the necessary parameters. POSLDA can also easily reduce to the nHMMLDA model by forcing all words in a sentence to share the same topic.

### 3.2 Approximate Inference

The principal computational problem in probabilistic topic/syntax models is posterior inference (Blei and Lafferty 2009). As it is based on LDA and the Bayesian HMM, exact inference in the POSLDA model is also intractable. Therefore, following many others, we make use of the MCMC-based approximate inference technique collapsed Gibbs sampling (Griffiths and Steyvers 2004; Heinrich 2004). Here, the multinomial parameters are first integrated out and we directly sample the indexing latent variables  $c_i$  and  $z_i$ . In POSLDA, if the class is designated as *syntactic*, then it only depends on the class. We therefore introduce the counts  $n_{w_i}^{(c_i, z_i)}$  which correspond to the number of times that word  $w_i$  is assigned to class  $c_i$  and topic  $z_i$ . Our sampling equation is then as follows:

$$p(c_i, z_i | \mathbf{c}_{-i}, \mathbf{z}_{-i}, \mathbf{w}) \propto \begin{cases} \rho_{c_i} \times \frac{n_{z_i}^{(d)} + \alpha_{z_i}}{n^{(d)} + \alpha} \frac{n_{w_i}^{(c_i, z_i)} + \beta}{n^{(c_i, z_i)} + W\beta} & c_i \in C_{\text{SEM}} \\ \rho_{c_i} \times \frac{n_{w_i}^{(c_i)} + \beta}{n^{(c_i)} + W\beta} & c_i \in C_{\text{SYN}} \end{cases} \quad (1)$$

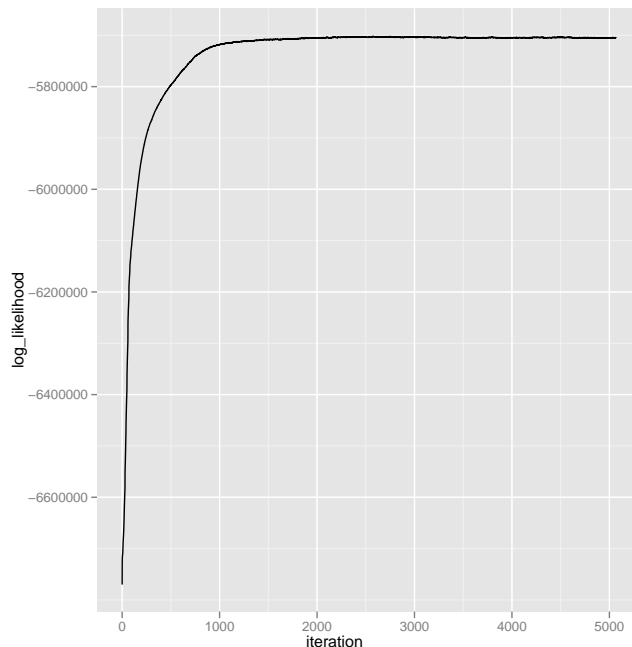
where

$$\rho_{c_i} = \frac{n_{(c_{i-2}, c_{i-1}, c_i)} + \gamma_{c_i}}{n_{(c_{i-2}, c_{i-1})} + \gamma} \cdot \frac{n_{(c_{i-1}, c_i, c_{i+1})} + \gamma_{c_i}}{n_{(c_{i-1}, c_i)} + \gamma} \cdot \frac{n_{(c_i, c_{i+1}, c_{i+2})} + \gamma_{c_i}}{n_{(c_i, c_{i+1})} + \gamma} \quad (2)$$

Note that we sample the pair  $(c_i, z_i)$  jointly as a block, which requires computing a sampling distribution over  $S_{\text{SYN}} + T \times S_{\text{SEM}}$ . It would also be valid to sample  $c_i$  and  $z_i$  separately, which would require only  $S + T$  computations, in which case, the sampling procedure would be somewhat different. Despite the lower number of computations per iteration, however, the sampler is likely to converge faster with our blocked approach because the two variables are tightly coupled. The intuition is that a non-block-based sampler could have difficulty escaping local optima because we are interested in the most probable *pair*; a highly probable class  $c$  sampled on its own, for example, could prevent the sampler from choosing a more likely pair  $(c', z)$ .

One problem with MCMC-based methods is that assessing convergence can be difficult. We do not address specific approaches to get around this issue, but we note that a stabilizing likelihood can be used to infer convergence. Generally this happens after several hundred iterations (depending on the size of the dataset), and for the datasets used in this article, the likelihood will converge at about 2,000 iterations. Finally, we are interested in deriving point estimates for the topics  $\phi^{(c, z)}$  from the sampled statistics.  $\phi$  is a 3-dimensional array where  $\phi_{w_i}^{(c, z)} = p(w_i | c, z)$ . Therefore, following (1), we get

$$p(w_i | c, z) = \frac{n_{w_i}^{(c, z)} + \beta}{n^{(c, z)} + W\beta} \quad (3)$$

**Figure 2**

Gibbs iterations vs. log-likelihood while learning a POSLDA model from the TREC AP dataset with  $K = 50$ ,  $S = 10$ ,  $S_{SEM} = 5$ .

## 4. Experiments and Results

In this section we present a set of experiments on the POSLDA model to demonstrate its capabilities as a topic and syntax model of language. We demonstrate both qualitatively and quantitatively the model’s ability to capture the semantic and syntactic axes of information prevalent in a corpus. We begin qualitatively with topic interpretability and then present quantitative results on the ability of POSLDA as a predictive language model. Following this, we show its ability as a model for performing unsupervised POS tagging.

### 4.1 Topic Interpretability

Judging the interpretability of a set of topics is highly subjective. Chang, et al. look at “word intrusion” where a user determines an *intruding* word from a set of words that does not thematically fit with the other words, and “topic intrusion” where a user determines whether the learned document-topic portion  $\theta_d$  appropriately describes the semantic theme of the document (Chang et al. 2009). Here, we are mostly interested in subjectively demonstrating the low incidence of “word intrusion” both in terms of semantics (theme) and syntax (part-of-speech). We subjectively demonstrate that our

<i>"law"</i>			<i>"finance"</i>			<i>"health"</i>		
<i>adj</i>	<i>verb</i>	<i>noun</i>	<i>adj</i>	<i>verb</i>	<i>noun</i>	<i>adj</i>	<i>verb</i>	<i>noun</i>
federal	filed	attorney	stock	rose	exchange	health	died	study
court	ruled	judge	wall	averaged	stock	medical	suffered	research
supreme	agreed	district	bond	issued	securities	aids	received	hospital
legal	contends	calif	million	fell	dow	drug	underwent	virus
civil	claims	county	american	gained	york	blood	found	report
appeals	contended	board	financial	dropped	inc	heart	carried	disease
tax	refused	loan	composite	rated	totaled	research	suffers	university
illegal	sued	san	common	traded	drexel	immune	leaves	doctor
government	won	court	business	stocks	commission	hospital	kills	person
financial	wrote	justice	dow	closed	lambert	cancer	took	patient

**Table 2**

Example topics learned from the TREC AP dataset with POSLDA.

model learns semantic and syntactic word distributions that are likely robust towards problems of word intrusion.<sup>2</sup>

To demonstrate the effectiveness of POSLDA’s semantic-syntactic pattern recognition ability, we fit a number of models to different datasets. We demonstrate results both on more “traditional” news corpora such as TREC AP and the WSJ, and on more esoteric datasets such as collections of tweets from the microblogging website *Twitter* and collections of legal decisions from the Supreme Court of Canada.<sup>3</sup> We begin with a standard demonstration of topic interpretability on news data from the Associated Press.

**4.1.1 Traditional News Data.** Table 2 shows three topics – manually labeled as “law”, “finance”, and “health” – learned from a 2,250 document subset of the TREC AP corpus (Harman 1992). We set the number of topics  $K = 30$ , the number of classes  $S = 17$ , and the number of semantic classes  $S_{\text{SEM}} = 7$ .<sup>4</sup> We show the top ten words from three POS-specific topics labeled manually as *adjective*, *verb*, and *noun*. The interpretability of the topics and the cohesiveness of the terms with high probabilities is clear. All three topics assign high probabilities to words that one would expect to have high importance. More importantly, however, the POS-specific topics also clearly reflect their syntactic roles. Each of the verbs is assuredly (even without the proper context) a verb, and the same thing for the nouns. The adjectives seem to fit as well; though many of the words could be considered nouns depending on the context, it is clear how given the topic each of the words can very well act as an adjective. A final point worth mentioning is that, unlike LDA, we do not perform stop-word removal. Instead, the POSLDA model has pushed stop-words to their own *syntactic* classes (rather than semantic) freeing us from having to perform pre- or post-processing steps to ensure interpretable topics. The top words in four of these topic-independent syntactic classes are shown in Table 3 with manually-labeled class names.

<sup>2</sup> This is in line with our approach of viewing topic models as *tools* for performing other tasks. We are most interested in objective quantitative results learned from applying our models to NLP tasks such as POS tagging which we demonstrate later in this section.

<sup>3</sup> <http://scc.lexum.org/en/index.html>.

<sup>4</sup> We choose this number based on intuition. We imagine that of the 17 often-delineated parts-of-speech (Goldwater and Griffiths 2007), 6 or 7 will generally be theme-specific. These include adjectives, verbs, gerund or present participle verbs, adverbs, nouns, and past participle verbs. It can be helpful to include an “extra” semantic class to see if the model can find patterns of semantics-syntax that we might miss.

AUXILIARY	CONJUNCTION	DETERMINER	RELATIVE
is	and	the	that
was	but	a	which
be	or	an	who
are	&	this	when
has	so	some	what
have	both	such	how
will	times	any	where
would	nor	many	whose
says	plus	those	why
were	yet	these	whom

**Table 3**

Example topic-independent syntax class distributions ( $\mathcal{C}_{\text{SYN}}$ ) learned from the AP dataset with POSLDA.

<i>“energy”</i>			<i>“corporations”</i>		
<i>adj</i>	<i>verb</i>	<i>noun</i>	<i>adj</i>	<i>verb</i>	<i>noun</i>
chemical	clean	plant	executive	named	president
power	waste	plants	vice	holding	company
environmental	attack	agency	operating	banking	officer
water	adore	utility	financial	managing	chairman
electric	exist	facility	management	succeed	board
energy	create	commission	company	succeeds	directory
air	combust	co.	board	named	post
pont	protect	industry	former	reacquired	unit
safety	mine	environment	investment	created	executive
gas	insult	corp.	division	centers	executives

**Table 4**

Example topics learned from the ACL\_DCI WSJ dataset with POSLDA.

Next we look at the ACL\_DCI release of the WSJ treebank dataset which contains approximately 3 million words over 6,058 documents. We turn to this dataset both to show further results of POSLDA’s pattern recognition ability along the axes of both semantics and syntax, and to demonstrate the scalability of the model to a larger corpus. Because this is a much larger dataset than the TREC AP corpus, we set the number of topics  $K = 50$ , but leave the class parameters untouched. Note that this approach to “guessing” the number of topics represented by a dataset is a typical way to begin understanding the make-up of a collection of documents. With the parametric version of POSLDA, we can use perplexity as calculated on a held-out test set to help determine the best number of topics and this is explored in the following subsection. For a more principled approach, we can use an HDP prior over the number of topics (Teh et al. 2006).

Table 4 shows two topics learned on the WSJ dataset with the parameters described above. Again, we show each topic as three POS-specific topics learned from POSLDA: verb, noun, and adjective. We show some of the more interpretable part-of-speech-specific topics, but in general the learned distributions appear noisier than those found on the smaller AP dataset. Nevertheless, the found topics are subjectively interpretable.

**4.1.2 Domain Specific Data.** While corpora of newswire documents are prevalent in NLP research due to their contained articles’ proclivity, there are a number of other

<i>"criminal"</i>			<i>"labour"</i>		
<i>adj</i>	<i>verb</i>	<i>noun</i>	<i>adj</i>	<i>verb</i>	<i>noun</i>
criminal	appeal	accused	labour	appeal	employer
mens	respect	offence	collective	finding	employee
bodily	committing	person	employment	issue	union
actus	section	code	union	respect	board
reasonable	determining	crown	bargaining	section	code
subjective	relating	act	trade	determining	agreement
mental	doing	conviction	individual	join	employees
objective	carrying	law	employee	colleague	court
unlawful	proof	defence	agricultural	lester	relationship
common	aiding	crime	construction	dismissing	position
<i>"insurance"</i>			<i>"family"</i>		
<i>adj</i>	<i>verb</i>	<i>noun</i>	<i>adj</i>	<i>verb</i>	<i>noun</i>
insurance	appeal	policy	spousal	appeal	court
insured	respect	insurer	child	account	spouse
hypothecary	effect	insured	family	determining	parties
disability	insure	contract	support	respect	agreement
unemployment	defend	clause	economic	regard	marriage
exclusion	issue	claim	financial	consider	act
life	accord	risk	matrimonial	considering	wife
automobile	indemnify	insurance	pension	accord	pension
third	force	premium	married	subsection	relationship
standard	insuring	beneficiary	marriage	section	husband

**Table 5**

Example topics learned from the Supreme Court of Canada decisions 1989 to 2009 with POSLDA.

domains with large collections of data that will benefit from new statistical models for corpus exploration, data mining, and other text-related tasks. These areas may include, *inter alia*, public health, economics, and law. Studying domain-specific corpora can be illuminating in text modeling research because often domains are filled with eccentricities that do not show up in general writing. These include domain-specific vocabularies and specialized writing structures. Law is a particularly relevant field because of the abundance of textual data that is produced in the field. Here, we demonstrate the qualitative effectiveness of modeling a collection of Supreme Court of Canada decisions with the POSLDA model.

We choose  $K = 40$ ,  $S = 17$ , and  $S_{\text{SEM}} = 7$ , as above. Four of the uncovered topics broken into adjective, verb, and noun are displayed in Table 5. Once again, the POS-specific topics are clear and interpretable. In the subtopic for "criminal law" adjectives, for example, there are a number of first-words from some common criminal law phrases. These include "mens" from the common phrase *mens rea* (the mental component required to commit a crime), "actus" from the common phrase *actus reus* (the physical act component required to commit a crime), and "reasonable" which often modifies the word "doubt" to form the phrase *reasonable doubt*. Furthermore, the nouns in the criminal law topic help make clear that domain specific data have been understood properly. The most probable word in the subtopic for "criminal law" nouns is "accused" which might naïvely be tagged as a verb in a dictionary-only based method

that has little understanding of context. Here, it is correctly placed in the *noun* subtopic because the *accused* is the person that stands accused of a crime.<sup>5</sup>

As with other datasets there is some noisiness in the results such as the word “section” appearing in the verb subtopic for the criminal, labour, and family law topics. This is likely due to the fact that in legal decisions phrases describing the origin of laws typically have their own sort of quasi-grammar and “section” is a common word in this context. An example is found in *Winters v. Legal Services Society*<sup>6</sup> where our sentence splitter found the sentence “The Requirements of Section 3(2)”. For this to be a proper sentence it requires a verb. The model has likely decided that the word “section” would work the best as a verb in this context and it is therefore likely an error attributable to the sentence splitting algorithm as opposed to the POSLDA model. Another problem is that the verb “appeal” has shown up as the most important verb in every topic. This is an interesting issue because typically we have seen indistinguishing words pushed to the syntax-only classes. One likely reason that this has not happened is that, though the word is important in many topics, it is not important in *all* of the uncovered latent topics. Appeal is also an interesting word in this domain because it is used to a great extent as both a verb (to *appeal* a decision) and as a noun (the *appeal* in question). While appeal is a common noun, it does *not* show up as any of the top ten nouns in any of the displayed noun subtopics. Despite these slight inconsistencies, however, the POSLDA-learned topics from the SCC dataset are interpretable and clear.

**4.1.3 Noisy Data.** Recently, there has been a large interest in mining the vast quantity of text data created every day though the popular microblogging service Twitter.<sup>7</sup> There are a number of unique challenges associated with analyzing this kind of data, however, that make it different from the datasets studied above. First, unlike the highly-structured news articles in corpora such as TREC AP and WSJ, Twitter messages (“tweets”) are rarely well-formed sentences. Proper grammar (or even anything close to it) is all but abandoned on Twitter and the rules of punctuation have been seemingly reinvented (Ramage, Dumais, and Liebling 2010). One of the principal reasons for this style of writing is that tweets are constrained to a maximum of 140 characters. This of course poses its own issues with respect to text modelling as short documents will contain less thematic information. In addition, partially due to character-limit constraints and partially because Twitter is very popular amongst young people, proper spelling is rare in such a dataset (Ritter, Cherry, and Dolan 2010). Each of these issues poses unique problems for modelling topics in this area. Because the long-range thematic dependencies in POSLDA are determined by word co-occurrence, multiple spellings of the same word can hinder unsupervised topic recognition. On the other hand, because the short-range syntactic dependencies in POSLDA are learned by understanding common word-class transitions, structureless grammar also causes problems in determining parts-of-speech.

Despite the issues outlined above, however, Twitter is a very interesting resource. It represents the up-to-the-minute thoughts of millions of people across the world and the knowledge that can be learned from this data is likely immense. One of the most interesting analyses of Twitter data so far is to use a supervised topic model to learn current issues about public health such as what health problems are being experienced

---

<sup>5</sup> See, e.g. Black’s Law Dictionary (2d Pocket ed. 2001).

<sup>6</sup> *Winters v. Legal Services Society*, [1999] 3 S.C.R. 160.

<sup>7</sup> <http://www.twitter.com>.

“media”		“relationships” (?)		“movies” (?)	
verb	noun	verb	noun	verb	noun
phone	photo	spent	girl	watch	online
posted	video	headed	party	looking	movie
speak	smile	cheating	sex	bored	script
looking	phone	visit	bitches	seen	series
send	night	annoy	sleep	walk	avatar
follow	time	callin	eyes	thats	funny
listen	media	wanna	love	respect	cool
chat	happy	hang	rings	cleared	bad
love	chick	f*ck	games	announce	girl
heart	da	sucks	date	wow	dollar

**Table 6**  
Example topics learned from Twitter with POSLDA.

by whom (and where) and how people are treating those problems (Paul and Dredze 2011). Here, we simply demonstrate that without any additional machinery, POSLDA can learn semantically and syntactically consistent topics from a collection of tweets. We use the Twitter POS dataset released at ACL 2011 by Gimpel, et al. which consists of approximately 26,000 words across 1,827 tweets (Gimpel et al. 2011). While this is a fairly limited collection of data, our model is nevertheless able to uncover some interesting POS-specific topics. Table 6 shows three manually-labeled topics learned with the settings of  $K = 20$ ,  $S = 17$ , and  $S_{SEM} = 7$ .

The topics outlined in Table 6 are by far the noisiest and least interpretable demonstrated thus far. Because of the non-standard text structure and issues with incorrectly spelled words, it is difficult for the algorithm to uncover the patterns of interest. Nevertheless, of the 20 topics, three fairly interpretable topics are demonstrated. As it is difficult to tell exactly which part-of-speech each subtopic represents, we only list two distinct subtopics per topic: verb and noun. The first topic – “media” – is likely the most coherent while the other two have been labelled with trailing question marks to convey that it is difficult to name the topics. It is for this reason that Twitter-specific topic models (or at least Twitter-specific alterations to existing topic models) may be required (Ramage, Dumais, and Liebling 2010; Ritter, Cherry, and Dolan 2010). Analyzing the listings in Table 6 does show, however, that POSLDA can uncover interesting semantic and syntactic patterns even in this highly noisy source.

## 4.2 Quantitative Results

There are several methods that are commonly employed to evaluate novel probabilistic models in the literature (Wallach et al. 2009). The original LDA paper – and many others – use the *perplexity* which is a standard metric in the information retrieval literature (Blei, Ng, and Jordan 2003; Teh et al. 2006). A probabilistic model can also be evaluated by considering its performance on an extrinsic task. Here, we first focus on the perplexity and use it to measure POSLDA’s performance as a predictive language model, and then discuss using the model for unsupervised POS tagging.

**4.2.1 Predictive Language Modelling.** Following the standard practice in topic modeling research (Blei, Ng, and Jordan 2003; Griffiths et al. 2005; Teh et al. 2006; Zhu, Blei, and Lafferty 2006), we fit a model to a training set and compute the perplexity of a held-out test set. The perplexity can be defined as the predicted average number of words that

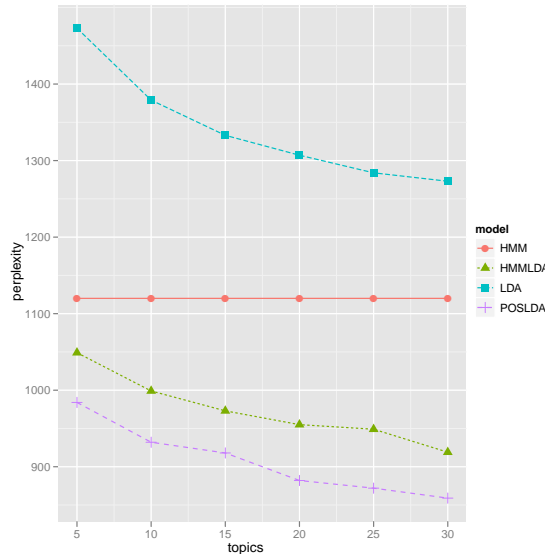
are equally likely to be generated for a given position (Zhu, Blei, and Lafferty 2006). It is also a monotonically decreasing function of the log likelihood.

The perplexity of a held-out test set  $\mathcal{D}_{test} = (\mathbf{w}_d)_{d=1}^M$  is given as:

$$ppx(\mathcal{D}_{test}) = \exp \left( - \frac{\sum_{d=1}^M \log p(\mathbf{w}_d | \mathcal{M})}{\sum_{d=1}^M N_d} \right) \quad (4)$$

where  $\mathcal{M}$  represents the model parameters learned from the training data,  $p(\mathbf{w}_d | \mathcal{M})$  is the probability (likelihood) of document  $\mathbf{w}_d$  given the learned model parameters, and  $N_d$  is the number of words in document  $\mathbf{w}_d$ .

We test the POSLDA model on a subset of the AP TREC dataset. We use ten-fold cross-validation where the data is split into 10 subsets of equal size. We conduct 10 experiments where one of the subsets is held out for testing and the model is trained on the remaining 9 subsets. We report the average over the 10 experiments. We compare the perplexity of POSLDA to the original LDA model, a Bayesian HMM, and Griffiths, et al.'s HMMLDA. Each model is trained using 10,000 iterations of Gibbs sampling. We use asymmetric priors on the document-topic portions  $\theta$  and the rows of the HMM transition matrix  $\pi$ . Following (Wallach, Mimno, and McCallum 2009), we use a symmetric prior on the topic-word distributions  $\phi$  because having asymmetric values on the topics themselves does not seem to lead to improved performance. The prior's hyperparameters are optimized using Minka's fixed-point method (Wallach 2008). For these experiments we use  $S = 10$  classes of which 5 are designated as semantic. By definition, the HMMLDA model has  $SS = 1$  and the LDA model has  $S = S_{SEM} = 1$ . The HMM model does not consider the number of topics  $K$ .

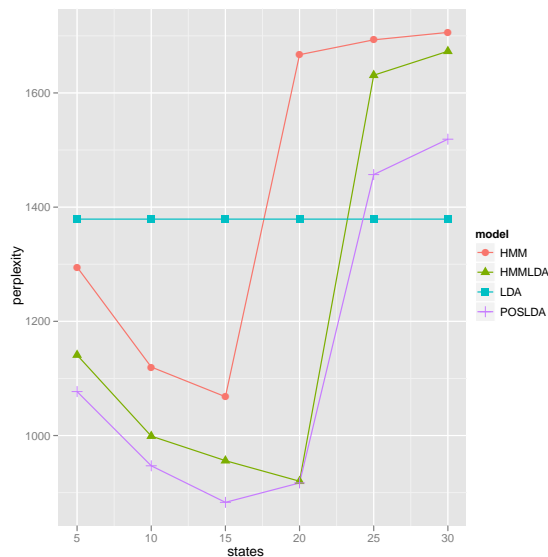


**Figure 3**  
Perplexity of POSLDA and other similar probabilistic models as  $K$  varies.

Figure 3 shows the average perplexity values on a held-out test set for a number of models in the same family as POSLDA over a range of topic values. The HMM achieves



lower perplexity values than LDA at all topic settings  $K = \{5, 10, 15, 20, 25, 30\}$ . All three topic-based models realize perplexity improvements as the number of topics  $K$  increases. Both HMMLDA and POSLDA – which combine the benefits of the HMM and LDA models – result in lower perplexity values. POSLDA’s additional flexibility in terms of the additional semantic classes allows it to record the lowest perplexity values of all the models tested for each topic setting.

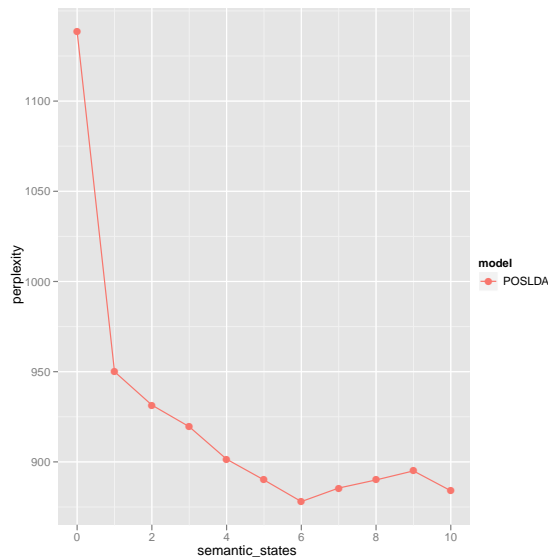


**Figure 4**  
Perplexity of POSLDA and other similar probabilistic models as  $S$  varies.

While Figure 3 shows how the POSLDA model’s ability to generalize on unseen data is affected by the number of topics, Figure 4 illustrates the changes in the perplexity when the number of classes  $S$  is the independent variable. While it may not reflect the best possible values for the POSLDA model, we set the number of semantic classes  $S_{\text{SEM}} = S$ . Interestingly, the HMM’s perplexity starts to shoot up when  $S = 15$ , and both HMMLDA and POSLDA show poor perplexity when  $S = 20$ . This likely reflects too much variation in the model (especially with no disambiguation between semantic and syntactic classes). Next, we look at how the perplexity varies when  $S$  is held fixed but  $S_{\text{SEM}}$  is free to vary.

Figure 5 shows the average perplexity as the number of semantic states  $S_{\text{SEM}}$  is varied. We set  $S = 10$  and investigate how different settings of  $S_{\text{SEM}}$  affect the model’s ability to generalize. When  $S_{\text{SEM}} = 0$  the model reduces to a Bayesian HMM and when  $S_{\text{SEM}} = 1$  it becomes the HMMLDA model. Adding some semantic information with HMMLDA improves the perplexity considerably (1172 to 957) and further distinguishing semantic information in POSLDA continues to improve the perplexity until it reaches a minimum at  $S_{\text{SEM}} = 6$ . This reflects the fact that certain classes of words such as conjunctions are not aided by thematic information.

Above we have demonstrated both that POSLDA tends to learn interpretable topics that are part-of-speech specific, and that it can lead to better predictive performance than other similar generative probabilistic models. The purposes of the models are clearly different, however. POSLDA is more flexible, but it requires a more expensive



**Figure 5**  
Perplexity of POSLDA as  $S_{SEM}$  varies.

representation of text (a sequence of words rather than a bag). HMMs are also extremely useful and versatile for a number of applications not necessarily related to text. Next we are interested in applications where the POSLDA model is truly needed or can truly make a difference. In (Darling and Song 2011) we showed how a POSLDA-like model can improve the results in a text summarization task. In the next section, we concentrate on unsupervised part-of-speech tagging.

### 4.3 Unsupervised POS Tagging

Goldwater and Griffiths show that Bayesian HMMs increase the accuracy of unsupervised POS tagging by up to 14 percentage points over the MLE approach (Goldwater and Griffiths 2007). While these results are impressive, unsupervised approaches continue to fall well short of the accuracy obtained with supervised taggers. Nevertheless, unsupervised approaches are preferred in many situations especially when there is no access to large quantities of training data in a specific domain or particular language. We therefore aim to continue improving accuracy with unsupervised approaches by introducing semantics as an additional source of information for this task.

The word “seal” appears both as a verb (to *seal* a jar or a leak) and as a noun (the marine mammal *Pinniped*) in the WSJ treebank dataset. The HMM approach can often tag each of these occurrences appropriately given the context, but there are cases where it will fail. However, if the topic being discussed is marine biology, we have another piece of evidence that increases the likelihood that this occurrence of the word “seal” is a noun about the marine mammal. If the topic is about pickling or roofing, however, “seal” as a verb is given more evidence. Another example is the word “book”. In a literary context it will almost always take the form of a noun. However, in a topic about promotions or services, the word is more likely to function as a verb: “to *book* a hotel”.

Following this intuition, we use the POSLDA model with the number of HMM states set to the number of possible tags in the tag set and use the state index learned through posterior inference as the predicted tag for each word.

We take two approaches to perform unsupervised POS tagging using the POSLDA model. The first approach uses a tag dictionary containing varying amounts of information on the possible tags that certain words have taken on in the training data. This renders the problem to a case of POS disambiguation rather than pure unsupervised tagging. It is, however, the most common approach to demonstrating results in unsupervised tagging (Goldwater and Griffiths 2007). The second approach is a pure unsupervised method that implements POS clustering. Because we cannot know which classes represent which parts-of-speech, we instead compare the learned clusters to the correct clustering where clusters are exchangeable. Accordingly, we use the variation of information (VI) for evaluating this task (Meilă 2007).

We showed in (Darling, Paul, and Song 2012) that this model consistently beats the Bayesian HMM approach in the Twitter domain. Here, we show that these improvements hold in a more traditional domain: the ACL\_DCI release of the Penn Treebank’s collection of Wall Street Journal newswire articles. This dataset contains approximately 3M words over 6K documents. We condense the tag set to the more standard for unsupervised POS tagging 17-tag set introduced by (Smith and Eisner 2005).

We follow the established form of (Merialdo 1993) and (Goldwater and Griffiths 2007) for unsupervised POS tagging by making use of a tag dictionary to constrain the possible tag choices for each word and therefore render the problem closer to disambiguation. Like in (Goldwater and Griffiths 2007), we employ a number of dictionaries with varying degrees of knowledge. A dictionary contains the tag information for a word only when it appears more than  $d$  times in the training corpus. We ran experiments for  $d = 1, 2, 3, 5, 10$ , and  $\infty$  where the problem becomes POS clustering. We report both tagging accuracy and the variation of information (VI), which computes the information lost in moving from one clustering  $C$  to another  $C'$ :  $VI(C, C') = H(C) + H(C') - 2I(C, C')$  (Meilă 2007). This can be interpreted as a measure of similarity between the clusterings, where a smaller value indicates higher similarity.

To properly make use of a tag dictionary, we slightly alter the POSLDA model by adding an additional observed random variable to the model as in Labeled LDA (Ramage et al. 2009). The tag dictionary is encoded in the model by a list of binary class indicators for each word  $\Lambda^{(w)} = (\lambda_{w_1}, \dots, \lambda_{w_V})$  where  $\lambda_w = (l_1, \dots, l_C)$  for each class. The model then becomes as depicted in Figure 1 (right). The sampling equation is then updated simply to  $p(z_i, c_i | \mathbf{c}_{-i}, \mathbf{z}_{-i}, \mathbf{w}) \propto p(z_i, c_i | \mathbf{c}_{-i}, \mathbf{z}_{-i}, \mathbf{w}) \times \lambda_{w, c_i}$ .

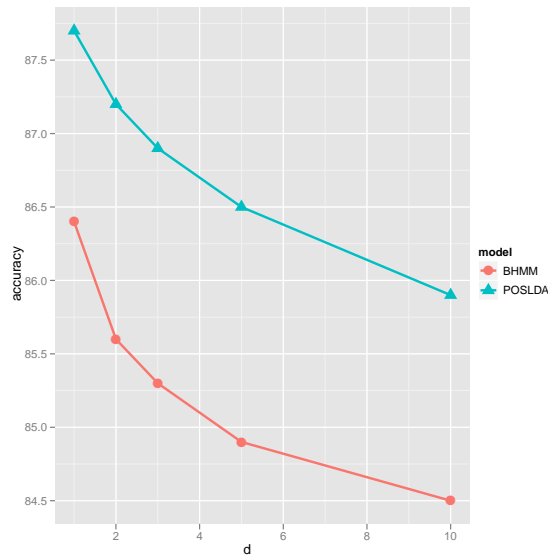
We run our Gibbs sampler for 20,000 iterations and obtain a maximum a posteriori (MAP) estimate for each word’s tag by employing simulated annealing. Each posterior probability  $p(c, z | \cdot)$  in the sampling distribution is raised to the power of  $\frac{1}{\tau}$  where  $\tau$  is a value (in traditional annealing used in physics  $\tau$  would be a temperature) that approaches 0 as the sampler converges. This approach is akin to bringing a system from an arbitrary state to one with the lowest energy thus viewing the Gibbs sampling procedure as a random search whose goal is to identify the MAP tag sequence, as employed in (Goldwater and Griffiths 2007). Finally, we run each experiment 5 times from random initializations and report the average accuracy and variation of information. All improved results reported for the POSLDA model are statistically significantly different from those achieved with the BHMM model as determined by a Student’s t-test with  $p \ll 0.01$ .

The achieved results are shown in Table 7 for a random tag assignment, the Bayesian HMM (BHMM) described in (Goldwater and Griffiths 2007), and our POSLDA tagging

Accuracy	1	2	3	5	10	$\infty$
random	58.7	57.9	57.5	56.7	55.5	
BHMM	86.4	85.6	85.3	84.9	84.5	
POSLDA	<b>87.7</b>	<b>87.2</b>	<b>86.9</b>	<b>86.5</b>	<b>85.9</b>	
VI						
random	2.37	2.44	2.48	2.54	2.63	5.07
BHMM	0.77	0.83	0.86	0.90	0.90	2.25
POSLDA	<b>0.73</b>	<b>0.77</b>	<b>0.79</b>	<b>0.82</b>	<b>0.86</b>	2.30
Corpus stats						
% ambig.	66.0	66.8	67.4	68.1	69.3	100
tags / token	2.34	2.48	2.57	2.71	2.95	17

**Table 7**  
POS Tagging Results on WSJ dataset.

approach. For both BHMM and POSLDA, we optimize the asymmetric hyperparameters  $\alpha$  and  $\gamma$  and the symmetric prior  $\beta$  using either direct optimization or sampling with similar results. We use 6 semantic classes – adjective, verb, gerund or present participle verb, adverb, noun, and past participle verb – leaving 11 remaining for syntax words, and we report results for  $K = 10$  topics.



**Figure 6**  
POS Tagging Accuracy on WSJ dataset.

These POS tagging results are also shown graphically in Figure 6. Our model outperforms the Bayesian HMM for all values of  $d$  in accuracy and all values of  $d$  but one for variation of information. In fact, POSLDA maintains higher tagging accuracy for  $d = 5$  than the Bayesian HMM for  $d = 1$ . Our approach consistently surpasses the results achieved with BHMM by approximately 1.5 percentage points for each value of

*d.* The one case where POSLDA did not improve upon BHMM is where  $d = \infty$  for POS clustering.

## 5. Conclusions and Future Work

In this article we presented the combined topic and syntax model, *Part-of-Speech* LDA or POSLDA. We have also demonstrated its use as an improved model for performing unsupervised POS tagging. Our overarching goal is to demonstrate that combining the two axes of word meaning – syntax and semantics – into a coherent model can result in improvements to both learned topic distributions and to NLP tasks such as POS tagging. We showed that incorporating semantic information into the HMM model led to improved results for this task. Additionally, we showed that combining the two axes of word information results in a language model that achieves lower perplexity – and therefore better predictive capability – than other similar probabilistic models.

In future work we would like to apply the POSLDA model to other NLP tasks that also rely upon learned word distributions. These include text summarization, text segmentation, and translation. An interesting avenue for further research is POSLDA's ability to serve as the base of a language generation system. While performing the LDA generative process would perhaps result in documents that contain words that are semantically related, we cannot say that it is generating language. POSLDA on the other hand – with some strong prior knowledge – may be able to generate coherent natural language because it exhibits more structure in the language generation process. We hope to explore this direction of research and apply it to tasks such as database population and abstract text summarization.

## Acknowledgments

The authors would like to thank the Natural Sciences and Engineering Research Council of Canada for partially funding this work through a Doctoral Post-Graduate Scholarship for William Darling. The authors would also like to acknowledge the financial support provided by Ontario Centres of Excellence (OCE) through the OCE/Precarn Alliance Program.

## References

- [Ahmed and Xing2010]Ahmed, Amr and Eric P. Xing. 2010. Staying informed: supervised and semi-supervised multi-view topical analysis of ideological perspective. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10*, pages 1140–1150, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Blei2012]Blei, David M. 2012. Probabilistic topic models. *Commun. ACM*, 55(4):77–84, April.
- [Blei, Griffiths, and Jordan2010]Blei, David M., Thomas L. Griffiths, and Michael I. Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *J. ACM*, 57(2).
- [Blei and Lafferty2009]Blei, David M and John D Lafferty. 2009. Topic models. *Text mining: classification, clustering, and applications*, 10:71.
- [Blei, Ng, and Jordan2003]Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.
- [Boyd-Graber and Blei2010]Boyd-Graber, Jordan L. and David M. Blei. 2010. Syntactic topic models. *CoRR*, abs/1002.4665.
- [Chang et al.2009]Chang, Jonathan, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*.
- [Chemudugunta, Smyth, and Steyvers2006]Chemudugunta, Chaitanya, Padhraic Smyth, and Mark Steyvers. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *NIPS*, pages 241–248.

- [Darling2012]Darling, William M. 2012. *Generalized Probabilistic Topic and Syntax Models for Natural Language Processing*. Ph.D. thesis, University of Guelph.
- [Darling, Paul, and Song2012]Darling, William M., Michael J. Paul, and Fei Song. 2012. Unsupervised part-of-speech tagging in noisy and esoteric domains with a syntactic-semantic bayesian hmm. In *Proceedings of the EACL 2012 Workshop on Semantic Analysis in Social Media*.
- [Darling and Song2011]Darling, William M. and Fei Song. 2011. Probabilistic document modeling for syntax removal in text summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT '11*, pages 642–647, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Gimpel et al.2011]Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- [Goldwater and Griffiths2007]Goldwater, Sharon and Tom Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751. Association for Computational Linguistics.
- [Griffiths and Steyvers2004]Griffiths, Thomas L. and Mark Steyvers. 2004. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235.
- [Griffiths et al.2005]Griffiths, Thomas L., Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2005. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press.
- [Harman1992]Harman, Donna. 1992. Overview of the first text retrieval conference (trec-1). In *TREC*, pages 1–20.
- [Hastie, Tibshirani, and Friedman2001]Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2001. *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. New York: Springer-Verlag.
- [Heinrich2004]Heinrich, Gregor. 2004. Parameter estimation for text analysis. Technical report, University of Leipzig.
- [Hofmann2001]Hofmann, Thomas. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196.
- [Jiang2009]Jiang, Jing. 2009. Modeling syntactic structures of topics with a nested hmm-lda. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining, ICDM '09*, pages 824–829, Washington, DC, USA. IEEE Computer Society.
- [Manning and Schütze1999]Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge (Mass.) and London.
- [Meilă2007]Meilă, M. 2007. Comparing clusteringsăĂŃan information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, May.
- [Merialdo1993]Merialdo, Bernard. 1993. Tagging english text with a probabilistic model. *Computational Linguistics*, 20:155–171.
- [Paul and Dredze2012]Paul, Michael and Mark Dredze. 2012. Factorial lda: Sparse multi-dimensional text models. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*. NIPS, pages 2591–2599.
- [Paul and Dredze2011]Paul, Michael J. and Mark Dredze. 2011. You are what you tweet: Analyzing twitter for public health. In *5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*.
- [Paul and Girju2010]Paul, Michael J. and Roxana Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *AAAI*.
- [Rabiner1990]Rabiner, Lawrence R. 1990. A tutorial on hidden markov models and selected applications in speech recognition. In Alex Waibel and Kai-Fu Lee, editors, *Readings in speech recognition*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pages 267–296.
- [Ramage, Dumais, and Liebling2010]Ramage, Daniel, Susan Dumais, and Dan Liebling. 2010. Characterizing microblogs with topic models. In *ICWSM*.
- [Ramage et al.2009]Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled lda: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1, EMNLP '09*, pages 248–256, Stroudsburg, PA, USA. Association for Computational Linguistics.

- [Ritter, Cherry, and Dolan2010]Ritter, Alan, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 172–180, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Salton and McGill1983]Salton, G. and M. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY, USA.
- [Smith and Eisner2005]Smith, Noah A. and Jason Eisner. 2005. Contrastive estimation: training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, pages 354–362, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Teh et al.2006]Teh, Y. W., M. I. Jordan, M. J. Beal, and D. M. Blei. 2006. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.
- [Wallach, Mimno, and McCallum2009]Wallach, Hanna, David Mimno, and Andrew McCallum. 2009. Rethinking lda: Why priors matter. In *NIPS*.
- [Wallach2008]Wallach, Hanna M. 2008. *Structured Topic Models for Language*. Ph.D. thesis, University of Cambridge.
- [Wallach et al.2009]Wallach, Hanna M., Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1105–1112, New York, NY, USA. ACM.
- [Wilson and Chew2010]Wilson, Andrew T. and Peter A. Chew. 2010. Term weighting schemes for latent dirichlet allocation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 465–473, Stroudsburg, PA, USA. Association for Computational Linguistics.
- [Wolpert and Macready1997]Wolpert, D. H. and W. G. Macready. 1997. No free lunch theorems for optimization. *Trans. Evol. Comp*, 1(1):67–82, April.
- [Zhu, Blei, and Lafferty2006]Zhu, Xiaojin, David M. Blei, and John Lafferty. 2006. Taglda: Bringing document structure knowledge into topic models. Technical Report TR-1553, University of Wisconsin, Madison.

—

—

—

—

—

—